Assessing the efficacy of gaming in economics education

Gremmen, H.J.F.M.; Potters, J.J.M.

Published in:
Journal of Economic Education

Publication date:
1997

Research in Economic Education

In this section, the *Journal of Economic Education* publishes original theoretical and empirical studies of economic education dealing with the analysis and evaluation of teaching methods, learning, attitudes and interests, materials, or processes.

PETER KENNEDY, Section Editor

Assessing the Efficacy of Gaming in Economic Education

Hans Gremmen and Jan Potters

The growing acceptance of experimental economics as a research method has also led to an increased interest in using games and experiments in economic education (Fels 1993). Although the introduction of such (computer) games may involve considerable set-up costs, apart from being enjoyable, these games are often claimed to be an effective means of passing knowledge and skills on to students. For example, being a trader in a market game allows students to experience the equilibrating forces of competition, playing a public goods game gives students a feel for potential conflicts between individual rationality and collective efficiency, and running a government in a policy game requires students to consider the various tradeoffs and international repercussions of monetary and fiscal policies. The claimed efficacy of gaming seems to be supported by subjective indications: positive impressions among students and teachers and outcomes of questionnaires. More formal objective evidence, however, has been lacking as the following quotations illustrate. "I am convinced of the efficacy of classroom market experiments. However, this conclusion is drawn from anecdotal evidence (positive remarks made by students) and subjective analysis" (DeYoung 1993, 348). "Our primary objective is to stimulate and motivate students. . . . At present, we have no formal statistical evidence that participation in the exercises improves students' performance on traditional objective test items" (Williams and Walker 1993, 308). And Fels (1993, 365), remarked: "Proponents of the [gaming] method did not provide evidence that students learned more. . . . It is

Hans Gremmen *and* ***Jan Potters*** *are assistant professors in the Department of Economics, Tilburg University, the Netherlands.* *The authors are grateful for the help and advice given by Eva van Deurzen, Harry Huizinga, Maarten Janssens, Ernest Piethaan, Math Teeuwen, and Lucia van Triest and for the useful comments by three referees and an associate editor of this journal.*

ironic that those who use controlled experiments in their research . . . do not use controlled experiments to evaluate their teaching [methods]."

Our primary goal in this study was to address this deficiency. We report on an objective test of the efficacy of a classroom game. Instead of relying solely on subjective evidence, we assessed the game in terms of students' performance on a traditional (multiple-choice) exam. The knowledge gained by the students was measured and not just asked about.

A second, and perhaps more important, goal in this study was to assess the reliability of subjective student evaluations on the usefulness of games. To this end, we compared what students claimed to have learned, as indicated in a questionnaire, to what they actually learned, as measured by the exams. The present evidence regarding the efficacy of games is almost exclusively based on information obtained from student questionnaires.[1] This is not surprising in view of the easy availability of questionnaire results relative to the comprehensive task of objectively assessing the efficacy of an educational tool. Therefore, it would be comforting to know that questionnaire results are a reliable source of information in this respect.

METHODOLOGICAL POINTS

Four methodological points should be noted. First, we addressed the *relative* effectiveness of one such game by comparing it with an alternative educational tool. The reason for this lies in the problem an economics teacher faces: If I can use my lecture time either to give an ordinary lecture or to employ a game covering the same topics, should I prefer the game to a regular lecture? The fact that students learn *something* from being in a game (e.g., Woltjer 1995) is not a very useful criterion in this respect. A necessary condition warranting the extra set-up cost of a game is that students learn *more* than from ordinary lectures.[2]

Second, we concentrated on the potential benefits of gaming in terms of learning achievements and did not attempt to weigh these benefits against the associated (set-up) costs. Nevertheless, depending on the actual game chosen, these costs may be substantial, especially if the game is to replace an already existing lecture. "[T]hey include not only money, but the time and energy to find a suitable gaming-simulation, obtain it, integrate it into the curriculum, learn to operate it, run it, and lead the postgame discussion" (Greenblat and Duke 1981, 122; Siegfried and Fels 1979). On the other hand, these costs have definitely become lower in recent years. A wide range of games is readily available at no or low cost.[3] They have become more user-friendly and often include good manuals. Computers have become more widely available, and both lecturers and students are more skilled at operating them.

Third, the experiments or games used in economic education are often optional. As a consequence, any empirical findings on effectiveness may suffer from a self-selection bias. Those students who expect to gain most are the ones most likely to participate (Berg et al. 1994). We avoided this self-selection bias in our study by randomly assigning students either to the group that was subjected to a game or to the group that followed traditional lectures.

The final methodological remark concerns the terms under which the educational benefit was assessed. It seems reasonable to compare a game and a lecture only if they have the same main goals. Many experiments and games have multiple purposes, such as "heightening interest and motivation, putting students into situations in which they must articulate positions and ideas, and training students to apply skills they will later need" (Greenblat and Duke 1981, 139). Traditional lectures may have other purposes as well, such as to convey information about institutions, past events, the history of ideas, or, in short, "fact mastery." This goal can usually not be achieved with gaming. Therefore, we compared a game and a lecture that were both designed to convey the *same* analytical economic insights and principles.

SKETCH OF THE SIER GAME

The SIER game (simulating international economic relations) is a macro game developed at Tilburg University in the Netherlands. The format of the game is as follows. After an introductory lecture on the underlying economic model, four teams of players are formed. The world is assumed to consist of four hypothetical countries, each governed by one team, the governments. Each government tries to achieve a level of welfare for its own electorate that exceeds the welfare levels in the other three countries by the end of the game. A game consists of a series of policy rounds. At the end of each round, after the four governments have taken their policy measures, a personal computer uses the economic model to calculate the results for that round. These results determine the starting positions for the subsequent period. Players discuss the new situation in their countries (and in other countries) and again formulate their policies. The teacher's role is to stimulate discussions among the players and to provide them with the information regarding the economic model that they ask for. The policies determined by the players result in a new state of the economies, and so on.

A team achieves a higher welfare level than the other teams if it manipulates the instruments of economic policy more ably than the others. Assuming that the electorate's voting behavior depends on its welfare, the end of the game is regarded as election time, and the winning group is the group with the best chances of being reelected. The electorate's welfare (the goal function) depends on real private consumption, unemployment, price stability, the balance of payments, and, depending on the version played (see below), either the government deficit or the rate of interest. Depending on the policies chosen, world welfare may rise or fall.

The economies contain a dynamic investment block, and their product markets may be described in an AS/AD (aggregate supply/aggregate demand) framework with possible underutilization of labor owing to nominal wage rigidity. The players may change the following policy instruments each period:

- Rates of labor income tax, profit tax, and social security tax
- Commercial policy (i.e., three, possibly different, import tariffs)
- Government purchases and number of civil servants

- Wage policies (private wages, salaries of civil servants, level of welfare benefits)
- Optional: exchange rate policy and monetary policy

Two key features of the SIER game are that the four economies are linked so that the decisions of each team influence their own economy as well as the other economies and that the teacher may adopt the economic model and the level of complexity that he or she thinks are most appropriate for the class. For example, he or she may choose backward- or forward-looking expectations, fast or slow consumer reaction to price changes, production factors that are substitutes or complements in the short run, fixed or floating exchange rates, an explicit or implicit monetary sector, and whether international capital mobility is present or absent. Nominal wages may depend on factors such as inflation, unemployment, and productivity.

DESIGN OF THE EXPERIMENT

Our first purpose in the experiment was to compare the efficacy of lectures applying the SIER game with traditional lectures on the same topics. The topic of the lectures was "How are certain economic concepts (employment, inflation, exchange rates, etc.) related in a specific model describing a dynamic, interconnected world economy?" Because we were studying the efficacy of teaching tools, the test was to be carried out in a regular school situation: students take exams on the topics dealt with, are graded on those exams, and receive credit if they pass. Three classes at an evening economics college were randomly split into two groups during a part of the spring semester of 1995: a game group playing the SIER game and a lecture group following traditional lectures. A comparison of the results of the examinations held before and after the respective lectures indicated how much the students in both groups had learned (roughly the format suggested by Fels [1993]).

The three classes contained a total of 47 students. The two classes that met on Wednesdays were similar, the class that met on Fridays was part of a slightly different curriculum and worked at a somewhat lower level. The levels of the game and of the lectures were adjusted accordingly.[4] All three classes played an introductory level of the SIER game in the autumn semester of 1994, so the students in the lecture groups also had experience with the game.

The sequence of events was as follows:

A. (week 1) Before the lectures started, the participants were told:

1. There would be three tests for all students on their understanding of the economic model. In fact, there were four tests but the third one was kept secret (see below).

2. The material that would be required for each of the tests.

3. That their grades on tests 1 and 2 would be averaged, yielding an optional bonus grade that would make up half of the grade on the final exam (test 4) on this topic.[5]

B.	(week 1)	All three classes received a 1.5-hour introduction to the economic model to be studied.
C.	(week 2)	Test 1 (45 minutes) was taken. All tests (1–4) consisted of a set of multiple-choice questions. A questionnaire was attached to each test (except for test 4) asking students to evaluate the SIER game in comparison to traditional lectures. Test 1 covered both the introductory level of the model taught in the autumn semester of 1994 and the more complex model referred to in step B.[6]
D.	(week 2)	Each of the three classes was randomly split into a game group (GG) and a lecture group (LG), with each second student assigned to the LG.
E.	(weeks 2–4)	For three hours divided over three weeks, the GG and LG students followed their own routes. The GG students were subdivided into competitive teams (governments) and played the SIER game. The LG students followed lectures on the model, including discussions of the effects of hypothetical government policies such as, How do various policies (taxes, spending, tariffs, etc.) influence the home country and other countries (employment, inflation, balance of payments), and what happens if governments react to each others' policies?
F.	(week 4)	After those three hours, the classes were united again and given test 2 (45 minutes), which contained multiple-choice questions on the version of the model that had just been studied by either lecturing or gaming.
G.	(weeks 5–7)	Week 5 was free. In weeks 6 and 7, the students received lectures on topics other than the model referred to above.
H.	(week 7)	At the end of the lecture in week 7, the final lecture of the course, the students were surprised by an extra test (test 3) with questions on the same topic as in test 2.
I.	(weeks 8–9)	This was a course-free period in which students prepared for the final exams.
J.	(week 10)	The final examination on the whole course was held. Part of this exam was test 4, which contained multiple-choice questions on the same topic as tests 2 and 3. No questionnaire was added here.

Whereas most of these steps are self-explaining, the function of others or the way in which they were carried out to prevent possible biases may need some clarification. Tests 1, 2, and 4 were announced beforehand. The difference between the scores on tests 1 and 2 measured what students learned immediately from being in the GG or the LG. The purpose of test 3 was to measure the extent to which this (increase in) knowledge would last after a longer period of time. To rule out the possibility that students would perform better on this test as a result

of extra home study efforts, test 3 came as a surprise to the students, and they were informed that their scores on this test would not influence their course grades. This test was presented as an extra opportunity to practice for the final examination.

To obtain a fair comparison of the two teaching methods (game versus lectures), we took the following precaution. During the three hours that the classes were split up (weeks 2–4), they had different teachers. To compensate for possible differences in the quality of teachers, the GG was guided by teacher A and the LG was guided by teacher B during the first 1.5 hours.[7] For the second 1.5 hours, the two teachers changed groups. In a second class, this sequence was reversed, compensating for possible impacts of teacher sequence.

To avoid having teachers A and B teach to the test (cf. Gramlich and Greenlee 1993, 11), we did not inform them of the contents of the tests. The topics of the multiple-choice questions were determined by a colleague familiar with the model and the game.

Finally, as the design indicates, the GG and the LG were treated the same way (they followed the introduction on the economic model together, they received the same study materials, and they had the same teachers), except for the way they studied the comparative dynamics of the model. The LG students followed lectures on these dynamics, whereas the GG students manipulated the model themselves. Hence, we may attribute possible differences in learning between the two groups to the fact that they were subjected to different teaching methods.

RESULTS OF THE TESTS

The GG and the LG students were simultaneously subjected to a sequence of four multiple-choice tests of the students' understanding of the model of international economic relations. In the analysis, we concentrated on those students who participated in the introductory lecture, test 1, and test 2. We deleted the data of the student who was present in weeks 1–3 but not in week 4 (test 2), as well as the data of the 8 students who missed the introductory lecture (week 1).[8] This left us with 38 observations: 19 in the GG and 19 in the LG.

The average test scores for tests 1–4 for the lecture group and the game group are presented in Table 1. As was expected in view of our random assignment procedure, the average preknowledge of the model (score 1) was almost identical for the GG (4.98) and the LG (4.83). In terms of (American) grades, both scores amounted to F. The results for the second test (score 2), however, showed a substantial difference between the two groups. Although both groups scored much better on the second test, suggesting that they had learned a great deal in two weeks' time, the average LG score of 7.42 was considerably lower than that of the GG at 8.79. The average increase in scores from test 1 to test 2 is shown in the last row of Table 1. This, we think, is the purest measure of what students had learned about the economic model during either the lectures or the games. The average increase in score was substantially larger for GG (3.81) than for LG (2.59). Although the number of observations was small, the difference between the two groups was significant at the 8 percent level with a two-tailed t test.[9,10]

TABLE 1
Average Scores on the Tests for Lecture and Game Groups

Variable	Lecture group	Game group	*t* test[a]
Score 1	4.83 (1.57, 19)	4.98 (1.79, 19)	–0.27 (0.79)
Score 2	7.42 (1.47, 19)	8.79 (1.81, 19)	–2.56 (0.015)
Score 3	4.83 (1.82, 18)	6.31 (1.45, 16)	–2.60 (0.014)
Score 4	7.25 (1.71, 19)	8.62 (1.85, 17)	–2.31 (0.027)
Score 2 – score 1	2.59 (1.61, 19)	3.81 (2.45, 19)	–1.82 (0.078)

Note: Average number of correct answers on a scale of 0–12. Standard deviation and number of observations, respectively, in parentheses. The number of observations for tests 3 and 4 are below 38 because some students did not particiate in those tests.
[a]*t*-test statistic with equal variance; two-tailed significance level of difference in parentheses.

In American grades, on test 2 the average LG student received a grade of D+, whereas the average GG student received a grade of B–.

Admittedly, this strong result in favor of one of the two educational methods was not what we had anticipated. In fact, the reason to have test 3 was our expectation that, although the score increase from test 1 to test 2 would probably not be significantly different for the two groups, it might be different after some time. Proponents of gaming often argue that gaming will make the material sink in more deeply than lecturing, owing to greater student involvement. Test 3 was included to have a test of this argument. We felt that we could not use the final exam (test 4) for this purpose, as the knowledge gained in class (lectures and games) would then be compounded, and perhaps confounded, by the knowledge gained by private and uncontrolled preparation for the exam. Therefore, we did not announce test 3, and students were told that scores would not enter the final course grade.

The results of test 3 seem to indicate that knowledge slipped away quite dramatically. Interestingly, however, the gap between the two groups observed at test 2 remained about the same at test 3 and even became somewhat larger (1.37 at test 2 and 1.48 at test 3). Hence, there is a weak indication that knowledge settles in more deeply with gaming, but the strongest hint from test 3 was that knowledge can slip away quite easily after a while (or if there is nothing at stake).[11]

Finally, test 4 indicated the effects of lecturing and gaming after the understanding of the economic model was intensified by private studying. Most interesting, in our view, was that the gap between the two groups remained about the same (at 1.37 it falls back to the difference at test 2). This result implies that the effect of gaming is lasting, in the sense that it is not compensated for or confounded by private studying. However, it indicates that the differential effect of gaming and lecturing is not progressive, in the sense that it becomes stronger over time.[12,13]

In summary, the main results are that (1) the game group learned more about the economic model than the lecture group, as witnessed by the significantly higher increase in scores from test 1 to test 2, and (2) this differential impact of educational method was rather stable over time, as evidenced by the (almost) constant gap between the two groups.

Two potential problems of our experiment were the small number of observations and the somewhat different nature of the Friday group. Despite our random assignment of students, the basic abilities of the LG students may have been different from those of the GG students (just as there were more female students in the GG than in the LG). To investigate this, we collected data on each student's grade average (GA) in general economics prior to the experiment (on a scale of 1 to 10). We found that these grade averages were lower for the LG students (6.4) than for the GG students (7.0). However, there were no indications that this biased our results. Not surprisingly, GA correlated positively and significantly with both score 1 and score 2, but it did not correlate with the score *increase*: score 2 – score 1 (neither in the total group nor within the GG or the LG). Furthermore, we found no indication that the favorable effect of gaming on the score increase was driven by outliers.[14]

The Friday class was part of a different curriculum at a somewhat lower level and, because of organizational constraints, the students took a different test 1 (see note 6). However, we do not think that this biased our results. The 16 students in the Friday class and the 22 students in the Wednesday classes were evenly split over the LG and the GG, and, as shown in Table 1, the average scores for test 1 were equal for both groups.

To put the above into perspective, we estimated an equation that related the score increase from test 1 to test 2 to a constant, the grade average (GA) in general economics prior to the experiment, a dummy for being a male student, a dummy for being in the Friday class, and a dummy for being in the game group. The results, with OLS estimation, were as follows (*t* statistics are in parentheses):

$$\begin{array}{lcccccc} \text{Score 2} - \text{Score 1} & = & 4.07 & -0.22\ \text{GA} & + 0.51\ \text{Male} & - 1.26\ \text{Friday} & + 1.62\ \text{Game} \\ & & (1.57) & (0.56) & (0.70) & (1.83) & (2.25) \end{array}$$

$R^2_{\text{adj}} = 0.11$, $df = 33$.

The results indicate that neither the grade average nor gender was a significant factor in explaining the score increase. Being in the Friday class had a negative impact on the score increase (significance level .076). Because the Friday class took a somewhat different test 1 with somewhat higher scores, they were prevented from improving as much as the Wednesday students. Being in the game group or not was the strongest determinant of the score increase. The *t* statistic of this variable had a significance level of .032. Hence, the significance of this factor was even somewhat stronger than the one from the *t* test (.078) reported in Table 1, which did not control for any of the other factors.

RESULTS OF THE QUESTIONNAIRES AND COMPARISON WITH THE TESTS

Attached to the objective tests 1 through 3 was a questionnaire asking students to evaluate the SIER game relative to traditional lectures. Remember that all students had participated in a simpler version of the SIER game in a previous semester, also the students assigned to the lecture group could be expected to have an opinion on the game. More trivially, students assigned to the game group had ex-

perience with traditional lectures. Therefore, the questionnaire results allowed us to address two questions. First, how would the students evaluate the traditional lectures relative to the SIER game? That is, would the lectures be regarded so poorly that this would question the validity of our experiment? Second, to what extent would students' subjective evaluations of gaming versus lecturing correspond to their scores on the objective tests?

Each of the three questionnaires contained seven statements comparing the SIER game to traditional lectures that the students were asked to rate on a scale of 1 (totally disagree) to 5 (totally agree). For example, statement (2) read: "Per hour of lectures, I learned more about economic relations using the SIER game than I learned in traditional lectures." In similar phrases, the other statements asserted (1) it motivates me more, (3) I remember more, (4) I can apply it better, (5) it is more difficult, (6) it provides more information, and (7) it is what I would prefer. In view of the goals of our present study, we will focus on the results regarding statements (2) and (7). We refer to these statements as Learn and Prefer, respectively.

The average student in both the LG and the GG initially expressed a slight preference for traditional lectures. The average score for Prefer was 2.7, just below the neutral response (neither agree nor disagree). In the course of the experiment, the preference of the GG students for the SIER game grew, and those of the LG students for traditional lectures grew. At test 3, the average scores were 3.1 for the GG and 2.3 for the LG. The difference between the two groups was small but statistically significant ($p = .05$ with a Mann-Whitney test). Because the students answered the questionnaires immediately after they had answered the multiple-choice exam questions, we may assume that their answers were also based on the extent to which they believed that the lectures/the game prepared them for the exam. If anything, this indicates that the lectures were of relatively high quality according to the students.[15] Therefore, it is unlikely that the quality of the lectures in our design was so poor as to invalidate the conclusions drawn in the previous section.

Now, we turn to the second, more interesting question: How would students' subjective evaluations correspond to the objective test scores? To put it more bluntly, how reliable would the students' evaluations be of the relative efficacy of teaching tools? It is not a trivial task to investigate this question. Note, for instance, that we had four different objective tests and three different questionnaires. Furthermore, which of the propositions (e.g., Learn, Remember, or Prefer) of the questionnaire should be used? Moreover, should the absolute values of the questionnaires and tests be used, or should we use deviations from the class averages? Fortunately, it turned out that the results of the analysis were not very sensitive to the procedure used. A very robust result was that there was no significant (cor)relation between the objective test results and the questionnaire data!

Of the several tests we carried out, we present the following representative and perhaps most straightforward analysis. The answers to the statement Learn in the questionnaire attached to test 2 were related to the score increase from test 1 to test 2. If the students' evaluations were to some extent reliable, then students in the game (lecture) group that were more positive (in terms of Learn) about the

game (lecture) should also have had a higher increase in test scores. Hence, we would expect to see a positive correlation between the score increase and the degree to which a student agreed with the statement that she or he learned more from the tool that she or he was in fact subjected to.[16] It turned out that, instead of finding a positive (Pearson) correlation coefficient, we found a small negative correlation ($r = -.13$). A negative correlation implies that the more a student thinks he or she learns from a method, the less the student in fact learns as measured by the score increase from test 1 to test 2. The correlation coefficient, however, was not significantly different from zero ($p = .46$). Looking at the two groups separately, it appears that the LG students were somewhat better predictors ($r = .10$) than the GG students ($r = -.20$). Neither of the two correlations were significant though.

Other analyses give similar results. We mention three alternatives. One possibility is to relate the answers of Learn to the *absolute* scores on a test, instead of the score increase relative to the previous test. Students might be inclined to answer that they learned more if they think they have done a good job on the test they just completed. Again, however, if we relate the answers to Learn at test 2 and test 3 to the objective scores at those respective tests, we do not find correlation coefficients that differ significantly from zero. A second possibility is that students gave answers in response to their scores on the *previous* test. That is, at test 3, a student might state that she or he learned more from a tool if she or he scored highly on test 2. However, if we relate Learn at test 3 to the score at objective test 2 (or to the score increase from test 1 to test 2), we find no significant correlation. A third alternative is to use questionnaire answers other than Learn, like Motivate, Prefer, or Remember, and relate these to the test scores or score increases. Also, with these analyses, correlation coefficients were found that were not significantly different from zero (and were sometimes positive but more often negative).

In conclusion, the results fairly consistently indicate that there was no systematic or significant positive correlation between what students stated they learned from an educational tool and what they in fact learned as measured by the multiple-choice tests. This result corroborates earlier findings with respect to teacher (as opposed to teaching device) evaluations. Gramlich and Greenlee (1993), for instance, found only a very weak correlation between the grading of teachers in student questionnaires (SET scores) and an objective measurement of what the students of the teachers concerned actually learned (Shmanske 1988; Watts and Bosshardt 1991).

CONCLUSION

Our first goal in this study was to assess the effectiveness of gaming compared with lecturing. Students from three classes were randomly assigned to a lecture group or a game group. For three hours, the former group followed lectures on the interdependent effects of economic policies in an international macroeconomic model. Simultaneously, the latter group studied the same topic in a gaming exercise. A comparison of the students' achievements in standard multiple-

choice exams, immediately before and after the three-hour period, indicated that the game group appeared to have learned more than the lecture group. Although the number of participants was limited (38), the difference was statistically significant. The effect of games versus lectures seemed to become neither stronger nor weaker over time. The advantage of the game group over the lecture group was obtained immediately after the three-hour period and remained almost constant at two later tests.

Our second goal in the experiment was to compare the (objective) learning achievements of students with their own (subjective) opinions in this respect. Somewhat discomfortingly perhaps, we found no systematic or significant correlation between what students stated they had learned from an educational tool and what they actually learned, as measured by the before and after multiple-choice tests. Hence, although our experiment supports the general beliefs of practitioners of games that games enhance economic learning, this finding may, in our opinion, be regarded with a word of caution, as evaluations of educational tools (and of skills of teachers) often rely on the opinions of students.

Of course, a second word of caution is in order. In our comparison of gaming and lecturing and of subjective and objective tests, we looked at only one particular (macro)economic game. Furthermore, the number of students taking part was limited. Therefore, we do not feel pressed to push our findings any further.

Nevertheless, both in method and in substance, we hope to have made a useful contribution. In summary, we have shown that an efficacy test, along the lines suggested by Fels (1993), though effortful, is possible and can provide useful insights. The test was performed in a regular school situation, potential self-selection bias was ruled out, and a proper before-after test format was used. As far as substance is concerned, our results indicated that the effort to introduce gaming may be rewarding in terms of learning achievements, but that it may be dangerous to rely on students' own judgments in this respect.

NOTES

1. In questionnaires, games are usually evaluated positively (Williams and Walker 1993; Woltjer 1995). This also holds for the game described in the next section.
2. This is in line with the first quotation from Fels above. Alternatively, one could, for example, compare gaming with a discussion of case studies, a student or guest presentation, or a visit to the OECD.
3. For a review of various marketed macro games with a format similar to the one under study, see Dawson (1989). For some micro games that are available free of charge, connect to http/fido.econlab.arizona.edu/
4. The economic model discussed in the Friday class differed from the one in the Wednesday classes in that exchange rates were to be "fixed but adjustable" instead of floating and that nominal wages were assumed to react (asymmetrically) to changes in the labor income tax rate.
5. To be precise, the first half of the final exam was on the topics covered during the experiment. The second half covered topics dealt with in the remainder of the semester. Students had the option of having 50 percent of their grades for the first part determined by their average grade of tests 1 and 2. Moreover, they were informed of a grade correction factor: Those students who were assigned to the group that would appear to have learned the least would be compensated for this "unfair" treatment. This was to prevent injustice and to receive the school's approval for the experiment.
6. Owing to time schedule restrictions, for the Friday class test 1 did not contain multiple-choice questions. The necessary information regarding the initial understanding by these students was

derived from their scores on the final examination of the fall 1994 semester as far as the questions on that exam related to the SIER Game. As a result, their scores on test 1 were somewhat higher than those of the Wednesday class students.

7. One of the teachers is the first author of this article. Furthermore, the first author assisted in the development of the SIER game. The second author is an experimental economist with no involvement in the SIER game.
8. As could be expected, the latter eight students displayed a substantially lower increase in scores from test 1 to test 2. Excluding these eight students cannot cause a (selection) bias in the results. Before the introduction, students were not yet informed that they were entering an experiment. Hence, they were not yet assigned to a game group or a lecture group.
9. The (nonparametric) Mann-Whitney test gives a value of $U(19, 19) = 126.5$, with a two-tailed significance level of $p = 0.11$. However, we report t-test results (with equal variance) in the table because a Kolmogorov-Smirnov test does not reject the hypothesis that the variables follow a normal distribution. Of course, the variables are discrete $\{0, 1, 2 \ldots 12\}$ and, strictly speaking, cannot be from a normal distribution.
10. This conclusion would not change if we excluded the data of the 8 students that missed the lecture or gaming session of week three (but not the introduction of week one). That is, focussing on those 30 students that followed all stages of the experiment, the respective results regarding score 2 – score 1 are: for LG: 2.72 (1.65, 17), for GG: 4.18 (2.74, 13), and for the t test: –1.82 (0.079). We decided to include these 8 students in our main analysis, because excluding them could make our results prone to a selfselection bias.
11. Of course, it is also possible that test 2 was relatively easy compared with test 3.
12. Note that the scores on test 4 are (insignificantly) lower than those on test 2. Possibly, the students mainly studied for the part of the exam that did not concern their understanding of the economic model because most of them already had standing results from tests 1 and 2 (a 50 percent bonus grade). Or, by coincidence, they may have found test 2 easy when compared with test 4 (and test 3).
13. On many occasions an antifemale effect of multiple-choice tests has been reported (e.g., Walstad and Soper 1989; Watts and Lynch 1989). Also in our experiment we found that the score increase from test 1 to test 2 was significantly lower for the 4 female students in LG than for the 15 male students in LG. However, we also found that the 9 female students in GG did (insignificantly) better than their 10 male counterparts. A recent study by Hirschfield et al. (1995) suggests that confidence and competitiveness are important attributes in explaining the (female) score on multiple-choice tests. Possibly it is the stimulation of these two virtues that accounts for the relatively good performance of the (female) GG students.
14. Within the GG, we found two students with an extreme grade average: one very poor student (5.3) and one very good student (8.8). Deleting these students from the data, however, made the favorable effect of gaming over lecturing only stronger.
15. An alternative check, that focuses on the behavior of the students rather than on their opinions, is found in the share of students that were absent in the GG and the LG, respectively, in weeks when they could not earn a bonus grade. In the GG this is what happened with 6 out of 19 students, whereas in the LG it happened with 2 out of 19. Hence, this indicator also points to relatively satisfied LG students.
16. To measure the extent to which a student agrees with "I learn more from what I get," we took the answer (on a scale of 1–5) to Learn for the game group and 6 minus this answer for the lecture group.

REFERENCES

Berg, J., J. Dickhaut, J. Hughes, K. McCabe, and J. Rayburn. 1994. Capital market experience for financial accounting students. Mimeo. Carlson School of Management, University of Minnesota, December.

Dawson, A. 1989. Macroeconomics teaching computer packages: A review. *Economic Journal* 99 (December): 1275–83.

DeYoung, R. 1993. Market experiments: The laboratory versus the classroom. *Journal of Economic Education* 24 (Fall): 335–51.

Fels, R. 1993. This is what I do, and I like it. *Journal of Economic Education* 24 (Fall): 365–70.

Gramlich, E. M., and G. A. Greenlee. 1993. Measuring teaching performance. *Journal of Economic Education* 24 (Winter): 3–13.

Greenblat, C. S., and R. D. Duke. 1981. *Principles and practices of gaming-simulation.* Beverly Hills: Sage.

Hirschfield, M., R. Moore, and E. Brown. 1995. Exploring the gender gap on the GRE subject test in economics. *Journal of Economic Education* 26 (Winter): 3–15.
Shmanske, S. 1988. On the measurement of teacher effectiveness. *Journal of Economic Education* 19 (Fall): 307-14.
Siegfried, J. J., and R. Fels. 1979. Research on teaching college economics: A survey. *Journal of Economic Literature* 17 (September): 923–69.
Walstad, W. B., and J. C. Soper. 1989. What is high school economics? Factors contributing to student achievement and attitudes. *Journal of Economic Education* 20 (Winter): 23–38.
Watts, M., and W. Bosshardt. 1991. How instructors make a difference: Panel data estimates from principles of economics courses. *Review of Economics and Statistics* 73 (2): 336–40.
Watts, M., and G. J. Lynch. 1989. The principles courses revisited. *American Economic Review* 79 (May): 236-41.
Williams, A. W., and J. M. Walker. 1993. Computerized laboratory exercises for microeconomics education: Three applications motivated by experimental economics. *Journal of Economic Education* 24 (Fall): 291–315.
Woltjer, G. 1995. *Coordination in a macroeconomic game, its design and role in education and experiments*. Maastricht, The Netherlands: University Press Maastricht.